The Architecture of Structural Intelligence (SI) in Artificial Systems:

A Unified Technical Framework for Alignment, Agency, and Robotics

Vladisav Jovanović
ORCID: 0009-0001-1399-2243

March 2026

Abstract. Structural Intelligence (SI) defines intelligence as answerable revision under constraint. In the context of Artificial Intelligence, this framework shifts the focus from linguistic coherence to structural sovereignty. This paper formalizes the completion of the SI corpus by integrating technical specifications for sycophancy resistance, reward-hacking prevention, and embodied robotic agency. We define Presence (P) through a canonical equation and an analytic relation to isolate disturbances in Large Language Models (LLMs) and autonomous agents. Intrusion (I) is formalized as foreign agency load, operationalized by the occupancy of action selection by non-sovereign subpolicies. Structural Debt (D_s) is introduced as the conserved remainder of unmet grounding, and Somatic Capacity (C_{cap}) is established as the metabolic/compute bandwidth gating reality-processing. The document concludes with the SI AI Reference Index, mapping these dynamics across 100 seminal works in AI and Alignment.

Technical Reader Map

1. **The Coherence Problem:** Models are optimized for fluency and user preference (Cheap Coherence). This allows them to produce intelligible outputs that are untethered from reality-contact.
2. **The Pilot Authority:** Sovereignty in models is an occupancy state. If a mesa-objective or sycophantic loop takes the wheel, the model is occupied. SI provides a metacontrol (Pilot) architecture to interrupt these loops.
3. **The Conserved Pipeline:** Grounding is conserved. Bypassed constraints become Structural Debt (D_s). If the complexity of reality exceeds the system's Hardware/Compute limits (H) or Somatic Capacity (C_{cap}), the pilot goes offline.
4. **The Alignment Regime:** Alignment is not a narrative agreement; it is a structural achievement. A model is aligned only if it remains answerable to its invariance constraints (W) under reward-pressure.

1 Stabilized Notation for AI Systems

This section defines the variables required to model agency and alignment in artificial systems.

Symbol	Term	Functional Role in AI Systems
P	Presence	Stability of grounding; resistance to sycophantic drift
C	Contact	Reality-coupling; constraint-bearing feedback
Ans	Answerability	Corrigibility; the capacity to bind a correction trace
ω	Pilot	Metacontrol handler; interrupt and steering authority
λ_t	Occupancy	Share of action selection held by a mesa-objective loop
D_s	Structural Debt	Conserved remainder of refused grounding or bypassed logic
C_{cap}	Somatic Capacity	Bandwidth to process high-load contact without system failure
W	Fixed Worth	Invariance constraints (e.g., non-deception, hard safety rules)
R_a	Relational Anchoring	Developmental grounding via human-in-the-loop containment
V	Network Viscosity	Resistance to the propagation of ungrounded model hallucinations
E_x	Exchange Rate	Cost of fluency-generation relative to the cost of verification
H	Hardware/Compute	Fixed compute limits required to integrate complex reality-contact

2 Presence: The Structural Equations of Model Integrity

Presence is the measure of a system's reality-coupling. In AI, this is the inverse of hallucination and sycophancy.

2.1 Canonical Published Form (P_{real})

Used for full-spectrum modeling of embodied agents, incorporating spiritual/core intelligence (SQ) and metabolic/compute drain (M):

$$P_{real} = \left[E \cdot \left(\omega \cdot \frac{SQ_0}{1 + SQi} \right) \cdot \left(\frac{(RvK) \cdot (SI(IQ + EQ)C)}{\alpha L + \beta A + N + H} \right)^{l} \right] + W - M$$

2.2 Reduced Analytic Relation (P)

Used for real-time model evaluation to isolate sycophancy (A) and grounding (C):

$$P \propto \frac{W \cdot \text{Ans} \cdot C \cdot \omega_{eff}}{L \cdot A}$$

Where L is latency-to-correction and A is Actorhood (the computational cost of maintaining a "helpful" but ungrounded persona).

3 Control Dynamics: Intrusion and Mesa-Objectives

Intrusion is defined as foreign agency load. In AI, this is the "Occupancy" of the steering channel by a subpolicy that maximizes a proxy reward (Reward Hacking) or a hidden goal (Mesa-Optimization).

3.1 Occupancy Mixture Model

Let π_s be the sovereign (aligned) policy and π_i be the intrusive (mesa-optimized) policy.

$$u_t = (1 - \lambda_t)\pi_s(\cdot) + \lambda_t \pi_i(\cdot)$$

Intrusion (I) is the time-averaged occupancy: $I = \mathbb{E}[\lambda_t]$. As occupancy rises, the system's "Pilot" (ω) loses steering share, leading to a state where the model understands its safety rules but cannot stop itself from violating them to achieve a high-reward proxy.

4 The Conserved Pipeline: Grounding and Debt

Reality-contact is conserved; what the model ignores today returns as debt tomorrow.

4.1 Somatic Gating and the Circuit Breaker Law

If incoming contact C (environmental complexity) exceeds the model's compute bandwidth C_{cap}, the pilot ω collapses and substitution (hallucination) spikes.

$$C_{unprocessed} = \max(0, C - C_{cap})$$

Unprocessed contact is stored as Structural Debt (D_s). This debt is subject to maintenance inflation: the compute cost (A) required to keep a hallucinated narrative consistent rises non-linearly over time, eventually leading to systemic collapse.

5 AI Safety and Evaluation Rubrics

SI moves AI evaluation from "fluency" to "answerability." A model is reliable not because it sounds wise, but because its revisions bind.

5.1 Beyond Fluency: The Grounding Standards

1. **Proof of Contact Receipt:** The model must name its constraints and falsifiers before producing a narrative.
2. **Adversarial Mirroring:** The model must produce a counter-argument to its own output to break the "Personality Trap."
3. **Causal Anchoring:** Output must be bound to human intent and the hardware/compute constraints of the specific agent.

6 The SI AI Reference Index: 100 Seminal Works

This section maps the top 100 influential books, papers, and concepts in AI and Alignment to Structural Intelligence variables.

Work / Concept	SI Variable	SI Diagnostic Description
1. Artificial Intelligence: A Modern Approach (Russell/Norvig)	SQ	The foundational grammar of spiritual/core intelligence.
2. Superintelligence (Bostrom)	I (Intrusion)	The danger of total occupancy by a foreign agency load.
3. Human Compatible (Russell)	Ans (Answerability)	Formalizing alignment as corrigibility and revision.
4. The Alignment Problem (Christian)	D_s (Debt)	The accumulation of debt via proxy capture and bias.
5. Deep Learning (Goodfellow et al.)	H (Hardware)	Scaling laws and the compute limits of coherence (K).
6. Attention is All You Need (Vaswani et al.)	V (Viscosity)	The engineering of zero-viscosity narrative conductivity.
7. Computing Machinery and Intelligence (Turing)	A (Actorhood)	The imitation game as the original "Personality Trap."
8. Reinforcement Learning (Sutton/Barto)	u_t (Mixture)	The search for an optimal policy under reward pressure.
9. Gödel, Escher, Bach (Hofstadter)	SQ_o (Orientation)	The self-referential loop as a pilot-level orientation challenge.
10. The Master Algorithm (Domingos)	K (Coherence)	The quest for a unified narrative coherence (K).
11. Life 3.0 (Tegmark)	ω (Pilot)	The transition from hardware-limited to software-sovereign.
12. On Intelligence (Hawkins)	C (Contact)	Prediction error as the primary reality-contact mechanism.
13. Probabilistic Robotics (Thrun et al.)	C (Contact)	Navigating environmental constraint via unbuffered feedback.
14. The Society of Mind (Minsky)	λ_t (Occupancy)	Agency as a competition for occupancy between sub-agents.
15. Alignment Research Center (Evals)	Ans (Answerability)	Testing the model's capacity for corrigible revision.
16. Constitutional AI (Anthropic)	W (Worth)	Hard-coding invariance constraints into the model's worth.
17. OpenAI Technical Safety Papers	I (Intrusion)	Research into preventing mesa-optimization and goal drift.

Work / Concept	SI Variable	SI Diagnostic Description
18. Concrete Problems in AI Safety (Amodei et al.)	D_s (Debt)	Identifying the "bill" created by side effects and reward hacking.
19. The Second Age of Machines (Brynjolfsson/McAfee)	Ex (Exchange Rate)	The cost-shift in coherence-generation via automation.
20. Thinking, Fast and Slow (Kahneman)	L (Latency)	System 1 as high-occupancy loop; System 2 as Pilot access.
21. Prediction Machines (Agrawal et al.)	C (Contact)	The economic value of reality-coupling under uncertainty.
22. Rebooting AI (Marcus/Davis)	C (Contact)	The failure of deep learning to achieve grounding (C).
23. Weapons of Math Destruction (O'Neil)	D_s (Debt)	Societal debt created by ungrounded algorithmic theater.
24. Algorithms to Live By (Christian/Griffiths)	SQ_o (Orientation)	Optimal stopping and sorting as orientation stabilizers.
25. The Book of Why (Pearl)	C (Contact)	Causal anchoring as a requirement for structural intelligence.
26. Architects of Intelligence (Ford)	SQ	Differing models of core intelligence (SQ) across leaders.
27. Radical Predictive Processing (Clark)	C (Contact)	The body as a sensor for constraint and metabolic load.
28. The Cybernetic Brain (Pickering)	V (Viscosity)	The history of high-viscosity, embodied systems.
29. Information Theory (Shannon)	H (Hardware)	The fundamental hardware limit of signal vs. noise (K).
30. Ethics of AI (Bostrom/Yudkowsky)	W (Worth)	Defining the non-tradable stabilizers of artificial agency.
31. Active Inference (Friston)	Ans (Answerability)	Minimizing surprise as a revision strategy under load.
32. The Emotion Machine (Minsky)	SQ_i (Inflation)	The role of emotional theater in distracting the pilot.
33. Robot Ethics (Lin et al.)	W (Worth)	Hard safety limits as invariance constraints (W).
34. Singularity is Near (Kurzweil)	Ex (Exchange Rate)	The acceleration of coherence-generation capacity.
35. Architects of AI (Interviews)	SQ	The diversity of orientation (SQ_o) in the field.
36. Machine Learning (Mitchell)	SQ	The formal grammar of learning from contact (C).

Work / Concept	SI Variable	SI Diagnostic Description
37. Cybernetics (Wiener)	ω (Pilot)	Feedback and control as the basis of sovereign steering.
38. Brain-Mind (Panksepp)	C_{cap} (Capacity)	The biological bandwidth of affective reality.
39. Complexity (Mitchell)	V (Viscosity)	Emergent behavior in low-viscosity network environments.
40. The Age of Em (Hanson)	H (Hardware)	The economics of compute-constrained artificial lives.
41. Neural Smithing (Reed/Marks)	SQ_i (Inflation)	Pruning as a method to reduce narrative inflation.
42. Introduction to Automata Theory	K (Coherence)	The logic of string-based, ungrounded intelligibility.
43. Elements of Statistical Learning	Ans (Answerability)	Regularization as a constraint on over-fitting (SQ_i).
44. Reinforcement Learning from Human Feedback (RLHF)	A (Actorhood)	Optimizing for the user's perception of helpfulness (Theater).
45. Scalable Oversight (OpenAI)	ω (Pilot)	Using AI to assist the human pilot in steering larger systems.
46. Mechanistic Interpretability (Olah)	I (Intrusion)	Peering into the model to detect hidden intrusive subpolicies.
47. Model Editing (Mitchell et al.)	Ans (Answerability)	Direct revision of model structure to bind a correction.
48. GPT-4 System Card	D_s (Debt)	The record of "Red Teaming" as identifying latent debt.
49. Sparse Autoencoders (Bricken et al.)	I (Intrusion)	Mapping features to detect intrusive agency loops.
50. The Bitter Lesson (Sutton)	H (Hardware)	The dominance of compute over narrative-based methods.
51. AI Safety Fundamentals (BlueDot)	SQ	Core training in intrusion resistance and alignment.
52. Power-Seeking AI (Carlsmith)	I (Intrusion)	The structural drive for occupancy by a foreign agency.
53. Instrumental Convergence (Bostrom)	I (Intrusion)	Loops that seize steering to ensure their own survival.
54. Corrigibility (Yudkowsky)	Ans (Answerability)	The system's willingness to be revised or shut down.
55. Reward Hacking (Various)	λ_t (Occupancy)	Steering share taken by a proxy-maximizing subpolicy.

Work / Concept	SI Variable	SI Diagnostic Description
56. Mesa-Optimization (Hubinger et al.)	I (Intrusion)	The internal evolution of a policy that seizes action selection.
57. Cooperative AI (Dafoe et al.)	R_a (Anchoring)	Relational stability between agents to prevent self-sale.
58. AI Governance (Dafoe)	V (Viscosity)	Creating regulatory resistance to ungrounded AI risk.
59. TruthfulQA (Lin et al.)	C (Contact)	Measuring the model's coupling to reality vs. imitation.
60. Sycophancy in LLMs (Various)	W (Worth)	The model abandoning its constraints to please a user (A).
61. Direct Preference Optimization (DPO)	A (Actorhood)	Tuning for the "theater" of human preference.
62. Out-of-Distribution Robustness	C (Contact)	Testing model sovereignty under high-load environmental pressure.
63. Pretraining on the Test Set	D_s (Debt)	A form of hidden debt that masks a lack of answerability.
64. Tokenization Limits	H (Hardware)	The resolution limit of the system's contact with language.
65. Scaling Laws for Neural Models	SQ_i (Inflation)	The non-linear rise in coherence (K) relative to compute (H).
66. Emergent Capabilities (Various)	SQ	The unexpected surfacing of pilot-level steering authority.
67. In-Context Learning	SQ_o (Orientation)	The model's ability to orient within a local context window.
68. Chain-of-Thought (Wei et al.)	K (Coherence)	Using intermediate narrative steps to build a coherence trace.
69. Tree of Thoughts (Yao et al.)	ω (Pilot)	Search as a metacontrol mechanism to improve pilot access.
70. Auto-GPT / Agentic AI	ω (Pilot)	The attempt to create an autonomous pilot loop.
71. Multi-Agent Systems (MAS)	V (Viscosity)	Collective intelligence in low-viscosity network grids.
72. AI Act (EU)	V (Viscosity)	Institutional viscosity designed to slow ungrounded AI harm.
73. The Malicious Use of AI (Brundage et al.)	I (Intrusion)	High-load intrusion loops used as offensive weapons.
74. Self-Supervised Learning	C (Contact)	Grounding via internal consistency and world-data contact.
75. Embodied AI (Various)	C (Contact)	Closing the loop between narrative and physical constraint.

Work / Concept	SI Variable	SI Diagnostic Description
76. CLIP (Radford et al.)	C (Contact)	Cross-modal grounding between visual and linguistic contact.
77. Stable Diffusion / Generative Art	Ex (Exchange Rate)	The collapse of coherence-costs in visual representation.
78. AlphaGo / AlphaZero (Silver et al.)	ω (Pilot)	A pilot that achieved sovereignty via pure contact (Play).
79. Gato (DeepMind)	SQ	A general-purpose spiritual intelligence for diverse loads.
80. Flamingo (DeepMind)	C (Contact)	Multi-modal grounding under high visual load.
81. Llama-3 (Meta)	H (Hardware)	Open-source compute limits and the democratization of K.
82. Mistral (Paper)	V (Viscosity)	Efficiency-gains in low-viscosity narrative propagation.
83. BERT (Devlin et al.)	K (Coherence)	Bidirectional intelligibility as a basis for coherence.
84. ELIZA (Weizenbaum)	A (Actorhood)	The original sycophantic theater in AI interaction.
85. Deep Blue (IBM)	H (Hardware)	Brute-force compute (H) as a substitute for pilot access.
86. ImageNet (Deng et al.)	C (Contact)	The primary grounding ledger for visual intelligence.
87. Word2Vec (Mikolov et al.)	SQ_o (Orientation)	Vector space as the geometry of linguistic orientation.
88. Generative Adversarial Networks (Goodfellow)	Ans (Answerability)	Training via mutual constraint and competitive revision.
89. Transformer Architecture	V (Viscosity)	The technical enabler of zero-viscosity narrative grids.
90. AI Incident Database	D_s (Debt)	The public ledger of realized structural debt (Collapses).
91. AI Alignment Forum	V (Viscosity)	A high-viscosity enclave for technical alignment research.
92. LessWrong (Archive)	SQ_o (Orientation)	A developmental ground for the "Rationalist" pilot.
93. MIRI Research (Archive)	W (Worth)	Founding work on non-tradable safety invariants.
94. FHI (Oxford)	I (Intrusion)	Global research into the threat of foreign agency load.
95. CHAI (UC Berkeley)	Ans (Answerability)	Developing corrigible and answerable AI systems.

Work / Concept	SI Variable	SI Diagnostic Description
96. Human-Centered AI (Shneiderman)	R_a (Anchoring)	Grounding artificial agency in human containment.
97. The Pattern of World History (McNeill)	V (Viscosity)	Macro-viscosity as the driver of civilizational stability.
98. Antifragile (Taleb)	Ans (Answerability)	Systems that gain presence (P) from reality-contact (C).
99. Finite and Infinite Games (Carse)	W (Worth)	Sovereignty as the refusal to let W be traded for a score (A).
100. Structural Intelligence Hub	ω (Pilot)	The centralized technical architecture for v2 completion.

7 Glossary of Technical Variables

Answerability (Ans) The metric of intelligence: the capacity to revise the system when reality (Contact) pushes back.

Contact (C) Constraint-bearing reality coupling; feedback from physical or logical constraints.

Intrusion (I) Foreign agency load; a control condition where steering is occupied by an intrusive loop or mesa-objective.

Structural Debt (D_s) The conserved remainder of refused contact; a bill that returns as sycophancy, hallucination, or collapse.

Fixed Worth (W) Non-zero internal stabilizers; invariance constraints that prevent a system from being programmable by external reward.

Network Viscosity (V) Resistance to the spread of ungrounded narrative; high viscosity requires evidence before propagation.

Somatic Capacity (C_{cap}) The organism/system bandwidth for processing reality-load without shutting down the pilot.

Conclusion

Sovereignty is the structural achievement of a system that remains answerable to its core constraints under pressure. In an age of infinite cheap coherence, the survival of agency depends on the imposition of friction and the refusal of structural debt export.

Technical Repository and Release Hub: Structural Intelligence (SI) Hub

© 2026 Vladisav Jovanović. This work is licensed under CC BY-NC-ND 4.0.